2025

AP® Statistics

Sample Student Responses and Scoring Commentary

Inside:

Free-Response Question 6

- ☑ **Scoring Guidelines**
- ☑ **Student Samples**
- ☑ **Scoring Commentary**

Question 6: Investigative Task — 4 points

General Scoring Notes

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4 (see the table at the end of the question).
- The model solution represents an ideal response to each part of the question, and the scoring criteria identify the specific components of the model solution that are used to determine the score.

	Model Solution	Scoring
A	Because the *p*-value of 0.002 is less than the level of significance of 0.05, the null hypothesis should be rejected. There is convincing statistical evidence of a difference between the mean reading score for all children, similar to those who participated in the study, who would read the story at 9 a.m. and the mean reading score for all children, similar to those who participated in the study, who would read the story at 3 p.m.	**Essentially correct (E)** if the response satisfies the following two components: 1. Provides correct comparison of the *p*-value to alpha (*p*-value is less than α) AND provides a correct decision about the null and/or alternative hypothesis 2. States a conclusion in context, consistent with, and in terms of the stated alternative hypothesis using nondefinitive language **Partially correct (P)** if the response satisfies only one of the two components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- To satisfy the *p*-value comparison in component 1, the response can compare the value of the test statistic to an appropriate critical value; for example, $|t| > 1.985$ if $df = 97.489$, or $|t| > 2.01$ if $df = 49$.
- An explicit decision about the null hypothesis is not required to satisfy component 1.
- If an explicit decision is stated and the conclusion is inconsistent with the decision, component 1 is not satisfied.
- The decision part of component 1 may be satisfied by implying the decision within the conclusion statement (sufficient evidence for the alternative hypothesis).
- To satisfy component 2, the response must include reference to means, groups (e.g., 9 a.m. and 3 p.m.), the sampling units (e.g., children), and the variable of interest (reading score).
- Examples of nondefinitive language in component 2 include “evidence to accept the alternative,” “there is evidence for the alternative,” and “there is not sufficient evidence for the alternative.”
- Examples of definitive language in component 2 include “accepts the null,” “proves the null,” “proves the alternative,” “accepts the alternative,” “there is not evidence for the alternative,” and “no evidence for the alternative.”

- If components 1 and/or 2 are satisfied and the response provides an incorrect interpretation of the p-value, the score is lowered from E to P or P to I.
- The quality of communication for responses with score P should be considered if holistic scoring is required.

	Model Solution	Scoring
B	It was appropriate for Stefan to conduct a two-sample *t*-test instead of a paired *t*-test because the two groups are independent. Stefan used random assignment to place the 100 volunteer children into two groups, and there is no indication that the two groups of 50 children are paired in any meaningful way (e.g., age, reading comprehension level).	**Essentially correct (E)** if the response satisfies the following two components: 1. States that the two-sample *t*-test is appropriate because the groups of children are independent *OR* that the groups of children were *not* paired in a meaningful way 2. Explains why the groups are independent instead of paired, either by referencing the use of random assignment of the 100 children to the two groups *OR* by providing an explanation that clearly separates the two situations, such as an example of how the data could have been paired **Partially correct (P)** if the response satisfies only one of the two components required for E *OR* if the response states that a two-sample *t*-test is appropriate because different samples were used but does not explicitly state that the children in the two samples were not paired in a meaningful way. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- For component 2 it is not sufficient to state that the reason the groups are independent is that one group was in the morning and one group was in the afternoon or because they received different treatments.
- For component 2 a sufficient explanation is one that provides an example of how the data could have been paired instead, such as if twins had been used or each child read a story at both 9 a.m. and 3 p.m.
- The quality of communication for responses with score P should be considered if holistic scoring is required.

	Model Solution	Scoring
C	i. The value of the pooled standard deviation is $s_p = \sqrt{\frac{(4.12)^2 + (4.43)^2}{2}} \approx 4.28$. Therefore, the value of Cohen's d is $d = \frac{\lvert 15.2 - 17.9 \rvert}{4.28} \approx 0.63$. ii. Based on Table 2, a Cohen's d value of 0.63 would indicate that Stefan's results were somewhat practical or meaningful in real life.	**Essentially correct (E)** if the response satisfies the following two components: 1. In part C (i) the response correctly calculates the value of Cohen's d with work shown. 2. In part C (ii) the response indicates the correct level of practical importance that is consistent with component 1. **Partially correct (P)** if the response satisfies only one of the two components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- An arithmetic or transcription error in a response can be ignored if correct work is shown.
- To satisfy the work-shown aspect in component 1, the response must show correct work for Cohen's d. Showing work for the pooled standard deviation is not required to satisfy component 1.
- A response that reports a negative Cohen's d and indicates the practical importance is "not very meaningful in real life" or indicates the practical importance is less than "not very meaningful in real life" may satisfy component 2.

	Model Solution	Scoring
D	i. The value of Cohen's *d* would decrease. If the standard deviation for the a.m. group and p.m. group were both greater than 4.43, the pooled standard deviation would be greater than 4.28. With a larger value in the denominator and the same value in the numerator, the value of Cohen's *d* would be smaller than 0.63. ii. The lower Cohen's *d* value would indicate less practical importance than that of the original results of Stefan's study.	**Essentially correct (E)** if the response satisfies the following three components: 1. In part D (i) the response states that the value of Cohen's *d* would be smaller. 2. In part D (i) the response provides a reason for the change in the value of Cohen's *d* that includes a correct reference to how Cohen's *d* will change based on an increase in the pooled standard deviation. 3. In part D (ii) the response states the decrease in Cohen's *d* would indicate there was less practical importance than the original study results. **Partially correct (P)** if the response satisfies only two of the three components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- If only the pooled standard deviation was calculated and no value is computed for Cohen's *d* in part C, the response may satisfy all components for part D if it explains that the larger standard deviations would result in an increase in the pooled standard deviation, and therefore show a lower effect size or less practical importance.
- A response that uses Table 2 and indicates that it is unknown if Cohen's *d* would be in the same or lower category of practical importance may satisfy component 3.
- The quality of communication for responses with score P should be considered if holistic scoring is required.

Scoring for Question 6

Each essentially correct (E) part counts as 1 point, and each partially correct (P) part counts as ½ point.

	Score
Complete Response	**4**
Substantial Response	**3**
Developing Response	**2**
Minimal Response	**1**

If a response is between two scores (for example, 2 ½ points), use a holistic approach to decide whether to score up or down, depending on the strength of the response and quality of the communication.

Question 6

Begin your response to **QUESTION 6** on this page.

A) Since $p = 0.002 < \alpha = 0.05$, we reject H_0. There's convincing evidence for a difference in the mean reading score of children between ~~to~~ reading at 9 am and reading at 3pm.

B) The children were not matched up to a similar child or twin where one child from the pair ~~is~~ reads in the morning and one reads in the afternoon so a matched pairs would not be appropriate. The two groups were independent of each other, making a two-sample t test appropriate.

C)

i) $s_p = \sqrt{\frac{4.12^2 + 4.43^2}{2}} = 4.2778$

$$d = \frac{|15.2 - 17.91|}{4.2778} = \boxed{0.631}$$

ii) 0.631 is between 0.2 and 0.8 so it would be interpreted as somewhat meaningful in life. This means there is a somewhat important difference in ~~the~~ the mean reading scores of ~~children~~ children between reading in the morning and reading at night.

GO ON TO THE NEXT PAGE.

Question 6

Continue your response to **QUESTION 6** on this page.

D)

i) ~~This would~~ The cohen's d coefficient in the new situation would be smaller than what was calculated in part C. This is because s_p would become greater as the standard deviations ~~would~~ increase because ~~the~~ numerator would get greater. A larger s_p means a bigger denominator in cohen's d coefficient formula resulting in a smaller answer because getting dinded by a bigger number decreases the answer.

ii) It would have less importance because Table 2 indicates that smaller cohen's d coefficients correspond to less importance. Part D(i) talks of a smaller coefficient meaning less importance.

GO ON TO THE NEXT PAGE.

Use a pencil or a pen with black or dark blue ink. Do NOT write your name. Do NOT write outside the box.

Question 6

Begin your response to **QUESTION 6** on this page.

A) Because the p-value of 0.002 is less than $\alpha = 0.05$, we reject H_0. There is convincing evidence that time of day does have an effect on reading comprehension in children

B) A paired t test is not appropriate because each group was assigned to only one of the treatments, not both.

C) i) $d = \frac{|15.2 - 11.9|}{2.068} = 1.306$ $\quad s_p = \sqrt{\frac{4.12 + 4.43}{2}} = 2.068$

ii) Stefan's results were 1.306 which is very meaningful. Based on Cohen's d coefficient value of 1.306, we have evidence that height can also be an indicator of reading scores

D) i) Cohen's d coefficient would be smaller because there is more variability which would lower its practical importance

ii) The new Cohen's d coefficient would have less practical importance than originally determined because there is more variability in the data.

GO ON TO THE NEXT PAGE.

Question 6

Begin your response to **QUESTION 6** on this page.

6.

A) We are 95% confident that we fail to reject the null hypothesis. We have convincing evidence because the p-value (.002) $< \alpha = .05$. The More 8th children in the study would read the story at 3:00pm than at 9:00am.

B) It was appropriate for Stefan to conduct a two-sample t-test for the difference in two populations means instead of a paired t-test for the pop. mean difference because of the two different times and some really read at both times or neither.

C)

i) $d = \dfrac{|\bar{x}_1 - \bar{x}_2|}{s_p}$ $\quad d = \dfrac{|15.2 - 17.9|}{s_p}$ $\quad = \dfrac{2.7}{4.278} \approx .6311$

$s_p = \sqrt{\dfrac{4.12^2 + 4.43^2}{2}}$

ii) The practical importance of Stefans results is that this study of when children are more likely to read 9am or 3pm is somewhat meaningful in life because $d = .6311 \rightarrow .20 < d < .80 \rightarrow .2 < .6311 < .80$ ✓

D)

i) The Cohens d coeffcient would be the same as the Cohens d coefficient calculated in part c. Because,

ii) the Cohen's d coefficient in part (d) indicate that Stefans observed difference in the means in the new situation would have less practical importance than what originally determined in part c(ii). Because,

Question 6

Note: Student samples are quoted verbatim and may contain spelling and grammatical errors.

Overview

NEW for 2025: The question overviews can be found in the *Chief Reader Report on Student Responses* on AP Central.

Sample: 6A
Score: 4

The response earned the following: part A – E, part B – E, part C – E, and part D – E.

In part A the response provides a correct comparison of the *p*-value to alpha and provides a correct decision about the null hypothesis, satisfying component 1. The response states a conclusion about the alternative hypothesis including reference to means, groups, the sampling units, and the variable of interest using nondefinitive language, satisfying component 2. This part of the response was scored essentially correct (E).

In part B the response identifies there are independent groups and explains why by stating, "The children were not matched up to a similar child or twin," satisfying both components. This part of the response was scored essentially correct (E).

In part C (i) the response correctly calculates the value of Cohen's *d* with work shown, satisfying component 1. In part C (ii) the response indicates "there is a somewhat important difference," satisfying component 2. This part of the response was scored essentially correct (E).

In part D (i) the response indicates that the value of Cohen's *d* would be smaller and provides a reason for the change based on the larger pooled standard deviation being in the denominator, satisfying components 1 and 2. The response does refer to an increase in the numerator; however, this is in reference to the pooled standard deviation and is therefore not an incorrect statement. In part D (ii) the response indicates there was less practical importance, satisfying component 3. This part of the response was scored essentially correct (E).

Sample: 6B
Score: 2

The response earned the following: part A – P, part B – P, part C – P, and part D – P.

In part A the response provides a correct comparison of the *p*-value to alpha and provides a correct decision about the null hypothesis, satisfying component 1. The response provides a conclusion about the alternative hypothesis, including reference to the sampling units and the variable of interest using nondefinitive language, but does not include reference to means or groups, so it does not satisfy component 2. For satisfying only one of the two components, this part of the response was scored partially correct (P).

In part B the response fails to state that the groups are independent or not paired, so it does not satisfy component 1. The response does not satisfy component 2 as it does not explain why the groups are independent instead of paired. However, because the response does indicate that different samples were used ("each group was assigned to only one of the treatments, not both"), the alternative requirement for partially correct is satisfied. This part of the response was scored partially correct (P).

In part C (i) the response does not satisfy component 1 as it does not correctly calculate the value of Cohen's *d*. The indicated work uses the standard deviations and not the variances in the calculation of the pooled standard deviation. In part C (ii) the response indicates the correct level of practical importance that is consistent with the Cohen's *d* in part C (i), satisfying component 2. For satisfying only one of the two components, this part of the response was scored partially correct (P).

Question 6 (continued)

In part D (i) the response indicates that the value of Cohen's *d* "would be smaller," satisfying component 1. The response does not refer to an increase in the denominator of the Cohen's *d* formula or refer to an increase in the pooled standard deviation leading to a lower effect size or less practical significance, so it does not satisfy component 2. In part D (ii) the response states that "The new Cohen's *d* coefficient would have less practical importance than originally determined," satisfying component 3. For satisfying components 1 and 3, this part of the response was scored partially correct (P).

Sample: 6C
Score: 1

The response earned the following: part A – I, part B – I, part C – E, and part D – I.

In part A the response provides a correct comparison of the *p*-value to alpha but does not provide a correct decision about the null hypothesis ("fail to reject the null hypothesis"), so it does not satisfy component 1. The response includes groups and the sampling units but does not include a conclusion about the alternative hypothesis using nondefinitive language, reference to means, and the variable of interest, so it does not satisfy component 2. Because neither component is satisfied, this part of the response was scored incorrect (I).

In part B the response does not satisfy component 1 as it does not state that the groups are independent or not paired, and the response does not satisfy component 2 as it does not explain why the groups are independent instead of paired. Because neither component is satisfied, this part of the response was scored incorrect (I).

In part C (i) the response correctly calculates the value of Cohen's *d* with work shown, satisfying component 1. In part C (ii) the response indicates that the level of practical importance is "somewhat meaningful in life," satisfying component 2. This part of the response was scored essentially correct (E).

In part D (i) the response does not satisfy components 1 and 2 as it states that "the Cohens *d* coefficient would be the same as the Cohens *d* coefficient calculated in part C," and it does not provide a reason for the decision. In part D (ii) the response states that "the new situation would have less practical importance," satisfying component 3. For satisfying only one of the three components, this part of the response was scored incorrect (I).